Econometría - ECN-323

Problemas de Especificación y de los Datos

Francisco A. Ramírez

Abril, 2020

Contenido

- Error de Especificación en la Forma Funcional
- Variables Proxy variables explicativas no observadas
- Modelos de pendientes aleatorias
- Propiedades MCO con Errores de Medición
- Datos Faltantes y Muestras no Aleatorias

Introducción

- Un supuesto de identificación del MRL es:

$$E(\varepsilon_i|x_i) = 0$$

- Es decir, que las x_i son variables exógenas.
- Cuando este supuesto no se cumple, se dice que x_i es una variable explicativa endógena.
- La implicancia es que los estimadores de MCO están segados:

$$E[b|x] = \beta + E\left[\frac{\sum(x_i - \bar{x})\varepsilon_i}{(x_i - \bar{x})^2}|x_i\right] \neq \beta$$

- Hay varias situaciones que pueden generar este resultado:
  - Omisión de una variable relevante del modelo.
  - Error de medición de una variable explicativa.
  - Diseño de muestras endógenas.
  - Forma funcional incorrecta.

Error de Especificación en la Forma Funcional

- En lo que sigue nos concentramos en el problema generado por forma funcional incorrecta.
- Suponga que el modelo verdadero es:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + beta_4 educ \times female + u$$

- El error de especificación en este caso podría darse porque:
  a. se omite la forma funcional correcta en que debe introducirse experiencia: $exper^2$
  b. se omite el término de interacción: $educ \times female$
  c. se usa $wage$ en lugar $log(wage)$

Prueba RESET para especificaci?n incorrecta de formas funcionales

- RESET: Prueba de Error de Especificación de la Regresión (RESET).
- Si el modelo correcto es:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

- entonces, al añadir formas funcionales de las variables explicativas ninguna será significativa.

- La prueba se basa en estimar:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error$$

- Con $H_o : \delta_1 = 0, \delta_2 = 0$, es decir, la forma funcional del modelo es la correcta.
- Se usa una prueba F de restricciones lineales con $n - k - 3$ grados de libertad.

Ejemplo: Ecuación para el precio de las viviendas

- Modelo 1:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

- Modelo 2:

$$log(price) = \beta_0 + \beta_1 log(lotsize) + \beta_2 log(sqrft) + \beta_3 bdrms + u$$

- donde: *price*: Precio de la vivienda, \$1000s *lotsize*: tamaño del lote (solar) en pies cuadrados *sqrft*: tamaño de la casa en pies cuadrados *bdrms*: número de habitaciones

Ejemplo: Ecuación para el precio de las viviendas

```
library(wooldridge)
library(lmtest)
library(stargazer)
data(hprice1)
attach(hprice1)
#Modelo lin-lin
mod.lin <-lm(price~lotsize+sqrft+bdrms)
#Modelo log-log
mod.log <-lm(lprice~llotsize+lsqrft+bdrms)
```

Table 1: Resultado Estimaciones

	Dependent variable:	
	price	lprice
	(1)	(2)
lotsize	0.002***	
	(0.001)	
sqrft	0.123***	
	(0.013)	
llotsize		0.168***
		(0.038)
lsqrft		0.700***
		(0.093)
bdrms	13.853	0.037
	(9.010)	(0.028)
Constant	−21.770	−1.297**
	(29.475)	(0.651)
Observations	88	88
R^2	0.672	0.643
Adjusted R^2	0.661	0.630
Residual Std. Error (df = 84)	59.833	0.185
F Statistic (df = 3; 84)	57.460***	50.424***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Ejemplo: Ecuación para el precio de las viviendas

```
library(lmtest)

#Modelo lin-lin
resettest(mod.lin,power=2:3)

##
##  RESET test
##
## data:  mod.lin
## RESET = 4.6682, df1 = 2, df2 = 82, p-value = 0.01202
#Modelo log - log
resettest(mod.log,power=2:3)

##
##  RESET test
##
## data:  mod.log
## RESET = 2.565, df1 = 2, df2 = 82, p-value = 0.08308
```

Observaciones sobre la Prueba RESET

- No provee indicación sobre cómo proceder si se rechaza el modelo.
- No se puede usar para problemas de especificación relativos a omisión de variables no observadas.
- Es solo una prueba para forma funcionales.

Variables proxy para variables explicativas no observadas

- Hay situaciones en las que no se observa una o un grupo de las variables explicativas.
- Estimar el modelo con las observables deriva en estimadores sesgados, si alguna de estas variables está correlacionada con la omitida.
- El uso de variables proxy de la omitida puede resolver o atenuar el problema.

Variables proxy para variables explicativas no observadas

- Sea el modelo verdadero:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

-Suponga que no se observa x_3^*, pero se tiene una proxy x_3 tal que:

$$x_3^* = \delta_0 + \delta_1 x_3 + v_3$$

- La 'solución al problema de variables omitidas' consiste en usar x_3 en el modelo en lugar de x_3^*.

- Para que esta estrategia funcione, es decir, proporcione estimadores consistentes de β_1 y β_2, tienen que cumplirse los siguientes supuestos:

  1. u no está correlacionado con x_1, x_2 y x_3^*.
  2. u no está correlacionada con x_3. Es decir, incluidas x_1,x_2 yx_3^* en el modelo x_3 es irrelvante.
  3. v_3 no está correlacionado con x_1, x_2 y x_3.

- El supuesto 3 es:

$$E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3) = \delta_0 + \delta_1 x_3$$

- Es decir, x_3^* tiene correlación cero con x_1 y x_2 una vez descontados los efectos parciales de x_3.

- Estos supuestos son suficientes para que esta estrategia funcione. Sustituyendo:

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3$$

- Defínase el error compuesto:

$$e = u + \delta_3 v_3$$

- Dado $E(u|x_1, x_2, x_3) = E(v_3|x_1, x_2, x_3) = 0$ entonces $E(e|x_1, x_2, x_3) = 0$.
- Así el modelo es:

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$

- Por lo que obtenemos estimados insesgados de β_1 y β_2. Así como de α_0 y α_3.

Table 2: Resultado Estimaciones

	Dependent variable:	
	lsalario	
	(1)	(2)
educ	0.065^{***}	0.054^{***}
	(0.006)	(0.007)
exper	0.014^{***}	0.014^{***}
	(0.003)	(0.003)
antig	0.012^{***}	0.011^{***}
	(0.002)	(0.002)
casado	0.199^{***}	0.200^{***}
	(0.039)	(0.039)
sur	-0.091^{***}	-0.080^{***}
	(0.026)	(0.026)
urbano	0.184^{***}	0.182^{***}
	(0.027)	(0.027)
afro	-0.188^{***}	-0.143^{***}
	(0.038)	(0.039)
IQ		0.004^{***}
		(0.001)
Constant	5.395^{***}	5.176^{***}

Modelos con pendientes aleatorias

- Hasta ahora el supuesto ha sido de que los coeficientes de las pendientes es el mismo para todos los individuos.
- O si varían lo hace en características medibles: sexo, estado civil, etc.
- Una especificación alternativa es el modelo de coeficiente aleatorio:

$$y_i = a_i + b_i x_i$$

- donde b_i se considera como una observación muestreada aleatoriamente de la población.
- Donde el modelo visto supone que $a_i = u_i$ y $b_i = \beta$.

- con una muestra de N datos, es imposible estimar este modelo, sino esperar estimar la pendiente e intercepto promedios.

$$\alpha = E(a_i)$$

y $\beta = E(b_i)$.

- es decir, β es el promedio del efecto parcial de x sobre y, es decir el *efecto parcial promedio (EPP)*

- Escribiendo

$$a_i = \alpha + c_i$$

$$b_i = \beta + d_i$$

- Por construcción $E(d_i) = E(c_i) = 0$. Sustituyendo

$$y_i = \alpha + \beta x_i + c_i + d_i x_i = \alpha + \beta x_i + u_i$$

- donde $u_i = c_i + d_i x_i$
- dados lo supuesto sobre c_i y d_i MCO producirá estimadores insesgados.

- No obstante, el error contendrá heterocedásticidad

$$var(u_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2$$

- a menos que $\sigma_d^2 = 0$, que en ese caso $b_i = \beta$ para toda i.

Propiedades de MCO bajo error de medici?n

- Existe la posibilidad que la variable económica de interés pueda ser medida con error.
- Este error de medición estará contenido en el modelo.
- A diferencia del caso de variable omitida, el interés puede ser sobre el coeficiente de la variable que contiene el error.
- Tipos de errores de medición
  1. en las variables dependientes
  2. en las variables explicativas

- Considere el siguiente modelo

$$y^* = \beta_0 + \beta_1 x_1 + u$$

- donde y^* es la variable de interés, se asume que hay una variable y con un ligero error de medición, definido como:

$$e_0 = y - y^*$$

- El modelo estimable es:

$$y = \beta_0 + \beta_1 x_1 + u + e_0$$

- Si e_0 no está correlacionado con x_i, MCO es insesgado y consistente.

- Lo que si es cierto, bajo esos supuestos

$$var(u + e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2.$$

Error de Medición en las Variables Explicativas

- Suponga el siguiente modelo:

$$y = \beta_0 + \beta_1 x^* + u$$

- donde se cumplen los supuestos de Gauss Markov, pero x^* es una variable no observda, con una medición con error.

$$e_1 = x_1 - x_1^*$$

.

- Se supone que en la poblacion $E(e_1) = 0$.

- Las propiedades de MCO dependen de los supuestos que se hagan sobre el error una vez se sustituya la variable del modelo por la variable medida con error:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

- este error compuesto, tiene media cero, por qué?

- la varianza es:

$$var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_e^2$$

- es decir, el error de medici?n hace que aumente la varianza del error.

- El supuesto de errores clásicos en las variables (ECV), dice que:

$$cov(x_1^*, e) = 0$$

- la medición observada es:

$$x_1 = x_1^* + e_1$$

- Asimismo,

$$cov(x_1 e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$$

- Respecto al modelo con el error compuesto:

$$cov(x_1, u - \beta_1 e_1) = -\beta cov(x_1 e_1) = -\beta \sigma^2_{e_1}$$

- Por tanto, en el caso de ECV, al regresi?n de MCO ofrece estrimados sesgados e inconsistentes.

- En particular,

$$plim(b_1) = \beta_1 + \frac{cov(x_1, u - \beta_1 e)}{var(x_1)}$$

$$= \beta_1 - \frac{\beta_1 \sigma^2_{e_1}}{var(x_1)}$$

$$= \beta_1 (1 - \frac{\sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}})$$

$$= \beta_1 (\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}}) < \beta_1$$

- A esto se le llama sesgo de atenuación.

Datos Faltantes

- En ocasiones los datos que disponemos padecen de datos faltantes en algunas de las variables clave del modelo.
- El efecto estadístico sobre la estimación depende de la razón de los datos faltantes:
    1. Datos faltantes al azar: la estimación es insesgada, aunque menos eficiente.
    2. Datos faltantes por muestreo no aleatorio

Muestras no aletorias

- En este caso significa que algunas observaciones tienen menos probabilidad de ser muestreadas.
- Tipos de selección muestral:
  1. selección muestral exógena: elección de la muestra en base a las variables exógenas. No causa sesgo en el estimador de MCO.
  2. selección muestral endógena: elección muestral basada en la variable dependiente. Causa sesgo en MCO.